METHOD AND COMPUTER PROGRAM PRODUCT FOR DRUG DISCOVERY USING WEIGHTED GRAND CANONICAL METROPOLIS MONTE CARLO SAMPLING

Inventors: Stephan Brunner and Charles Karney

This patent application claims the benefit of U.S. Provisional Patent Application 60/482,774 (filed June 27, 2003), U.S. Provisional Patent Application 60/509,272 (filed October 8, 2003), U.S. Provisional Patent Application 60/509,543 (filed October 9, 2003), and U.S. Provisional Patent Application entitled "Method and Computer Program Product for Drug Discovery Using Weighted Grand Canonical Metropolis Monte Carlo Sampling," serial number to be determined, SKGF Ref. 1866.0510000 (filed December 23, 2003), all of which are incorporated herein by reference in their entireties.

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The invention described herein relates to models for molecular interaction, and in particular the use of such models for drug discovery.

Related Art

[0002] In determining drug leads, it is often desirable to model a system that includes a protein and a set of small molecular fragments. Given the three dimensional structure of a target protein, usually obtained experimentally from x-ray crystallography, the basic interactions between the protein and the small fragments (typical average molecular weight of 150) are computed. This computation can be carried out by Monte Carlo (MC)-type modeling and analysis (usually implemented in software) for a large collection of organic fragments with diverse physico-chemical properties. The number of fragments can be in the hundreds to thousands. What are needed, therefore, are a method and computer program product for modeling such a system of fragments for purposes of determining drug leads.

SUMMARY OF THE INVENTION

[0003] The invention described herein includes a method and computer program product for modeling a system that comprises a protein and a plurality of fragments in order to identify drug leads. To analyze the interaction between a given fragment and a protein, the fragment states are sampled from a thermodynamically relevant Grand-Canonical distribution. The underlying sampling algorithm is a weighted Grand-Canonical Metropolis Monte Carlo approach, referred to herein as WGCMMC. The purpose of this weighted approach is to enable an essentially uniform numerical sampling of all states of interest of the fragment with respect to the protein, i.e. sampling deeper and shallower energy wells with the same thoroughness, while still avoiding the sampling of very unfavorable poses (e.g., as a result of steric clashes). The data is then finally re-weighted, so that the sampling correctly represents the considered thermodynamic ensemble. In practice, the weighting procedure is implemented by subdividing space with a grid. An orthogonal, equidistant grid is typically chosen. Each grid cell center x is assigned a local, numerical chemical potential field value $B_{num,}(x)$, which is adapted iteratively, based on previous sampling statistics, so as to ensure an approximately uniform numerical sampling of fragment states at all regions of interest around the protein. B_{num} is related to the energetic cost of inserting or removing a fragment from the numerical distribution, and the difference between its local value B_{num} (x) and the actual physical chemical potential B of the system defines the weight w for each sampled fragment state.

[0004] Once the B_{num} field has sufficiently converged, as a result of successive iterations, and the Markov chain associated with the Metropolis algorithm has equilibrated, the actual Monte Carlo sampling can be gathered. This is carried out by periodically saving the state of the system along the Markov chain. The number of Markov steps interspacing the gathered states must be sufficiently large to ensure proper decorrelation. Saving a state of the system involves the positions, orientations, potential energies and weights for all fragments currently present in the system. By making use of this fragment data, binding modes can then be identified and corresponding binding free energies estimated. The fact that the simulation system is considered in the framework of the grand canonical ensemble instead of the canonical ensemble enables, through simulated annealing of the chemical potential, efficient estimation of the free energy of binding of the fragment for various binding modes on the protein surface. This binding data for the different fragments can then in turn be used for identifying the relevant protein binding sites, and for assembling the different fragment types to obtain larger ligand molecules.

[0005] Further embodiments, features, and advantages of the present inventions, as well as the structure and operation of the various embodiments of the present invention, are described in detail below with reference to the accompanying drawings.

DESCRIPTION OF THE FIGURES

[0006] FIG. 1 is a flowchart illustrating overall processing of an embodiment of the invention.

[0007] FIG. 2 is a flowchart illustrating the initial step of preparing a molecular model for the system to be analyzed.

[0008] FIG. 3 is a flowchart illustrating the modeling process at the systemic level for computing the fragment-protein interactions using a weighted Grand Canonical Metropolis Monte Carlo approach, according to an embodiment of the invention.

[0009] FIG. 4 is a flowchart illustrating the convergence phase of the simulation system, according to an embodiment of the invention.

[0010] FIG. 5 is a flowchart illustrating the sampling phase of the simulation system, according to an embodiment of the invention.

[0011] FIG. 6 is a flowchart illustrating the process of identifying potential binding sites, according to an embodiment of the invention.

[0012] FIG. 7 is a flowchart illustrating the process of clumping fragments before assembly into drug leads, according to an embodiment of the invention.

[0013] FIG. 8 is a block diagram illustrating a computing platform on which a software embodiment of the invention can be stored and executed.

DETAILED DESCRIPTION OF THE INVENTION

[0014] A preferred embodiment of the present invention is now described with reference to the figures, where like reference numbers indicate identical or functionally similar elements. Also in the figures, the left-most digit of each reference number corresponds to the figure in which the reference number is first used. While specific configurations and arrangements are discussed, it should be understood that this is done for illustrative purposes only. A person skilled in the relevant art will recognize that other configurations and arrangements can be used without departing from the spirit and scope of the invention. It will be apparent to a person skilled in the relevant art that this invention can also be employed in a variety of other devices and applications.

I. Overview

[0015] The invention described herein is a fragment-based approach for designing drug leads. For this purpose, Locus Pharmaceuticals, Inc., Blue Bell, PA, developed the Locus Monte Carlo (LMC) code. The approach described herein makes use of a weighted Grand-Canonical Metropolis Monte Carlo algorithm for sampling fragments around the target protein. This sampling data can then be directly used for estimating the free energy of binding for different binding modes of the fragment on the protein surface. This approach distinguishes itself from a similar process implemented by Mezei and Guarnieri in their Metropolis Monte Carlo (MMC) code (Guarnieri, F. and Mezei, M., *J. Am. Chem. Soc. 118*:8493-8494 (1996)), in that it removes fragment-fragment interactions.

[0016] During the Monte Carlo sampling, a set of attributes are saved for each rigid fragment instance, including the coordinates of the fragment's center of mass (x,y,z), the quaternion $q = (q_1, q_2, q_3, q_4)$ characterizing its orientation, and the potential energy of interaction E between the fragment and the protein.

[0017] This LMC data for the different fragments can be analyzed for identifying potential binding sites using diagnostic tools such as the Locus Cluster Analysis (LCA) code and the Locus Binding Analysis (LBA) code (Locus Pharmaceuticals, Inc., Blue Bell, PA). These tools are based on the postulate that a binding site must be a localized high affinity region for a diverse collection of fragments, i.e. fragments with different physico-chemical properties. It is indeed assumed, that diverse interactions in a localized region are the necessary condition for ensuring the specificity of a binding site. If available, one naturally also makes use of experimental binding site data (e.g., co-crystal X-ray data and residue mutational analysis) in determining the final site within which the leads are designed.

[0018] Within the chosen binding site, fragments can be assembled into the actual candidate drug leads, usually composed of four to five fragments and thus having a molecular weight of the order of 600-800, using a software package such as the Locus Chemistry Design (LCD) software (Locus Pharmaceuticals, Inc., Blue Bell PA). Here again, use is made of the LMC fragment data in providing preferred fragment states -- positions and orientations -- with respect to the protein. Assembly of fragments is carried out based on geometric proximity, and using a variety of rules by which organic fragments may bond together. In somewhat more detail, two fragment states can be assembled, if the relative positions of their atoms enable, within given tolerances, to establish a certain type of bond, with specific bond lengths and angles. The most elementary bonding rule is of the form

$$-CH_2\text{-}H + H\text{-}CH_2\text{-} \quad \Rightarrow \quad \text{-}CH_2 - CH_2\text{ -} \quad .$$

[0019] Other bonding rules, such as the fusing of methyl groups or merging of cyclic rings may also be considered.

[0020] Fragment-based computational approaches are well-known. One example is the Multiple Copy Simultaneous Search (MCSS) numerical tool presently commercialized by Accelrys, of San Diego, California, and derived from an original version developed by the group of Karplus, Harvard University, MA, (Miranker, A. and Kaprlus, M., *Proteins: Struc. Func. Gen. 11*:29-34 (1991); Caflish, A., *et al., J. Med. Chem. 36*:2142-2167 (1993); Joseph-McCarthy, D., *et al., J. Am. Chem. Soc. 123*:12758-12769 (2001)). (These references are incorporated herein by reference in their entirety.)

[0021] What distinguishes the LMC approach from previous fragment-based methods is its ability to compute the actual thermodynamic fragment distributions around the protein, i.e. distributions consistent with thermal fluctuations at physiological temperatures. Information on the thermodynamic distribution is essential for computing free energies of binding, which, as presented further on, is the basic biologically relevant quantity for quantifying the binding affinity of a ligand.

[0022] Indeed, the MCSS approach for example is essentially based on an energy minimization procedure, providing fragment states corresponding to various local minima of the potential energy field representing the fragment-protein interaction. Such a procedure is computationally more expeditious than computing the actual physical, thermodynamic distributions, but is unable to provide information on entropic effects, essential for free energy estimates.

[0023] For computing the thermodynamic distributions, the LMC code package makes use of a Metropolis Monte Carlo approach (Metropolis, N., *et al., J. Chem. Physics 21*:1087-1092 (1953)) for sampling from a grand-canonical ensemble of states (Adams, D.J., *Molecular Physics 29*:307-311 (1975); Mezei, M., *Molecular Physics 61*:565-582 (1987)). (These references are incorporated herein by reference in their entirety.) In addition to exchanging just energy with a surrounding thermal bath, as in the case of a canonical ensemble, the system described by a grand-canonical ensemble exchanges particles (or fragments in the case of LMC) with its surroundings as well. The energy cost associated with inserting/deleting a fragment from the system is controlled by its chemical potential. By varying this chemical potential, so-called simulated annealing of the chemical potential, one may vary the average number of fragments in the simulation system. It is shown further on, that measuring the values of the chemical potential at which fragments leave various sites on the protein provides an estimate of the free energy of binding for the different binding modes over the protein surface.

[0024] The practicality of the simulated annealing procedure for estimating binding affinities was demonstrated by Guarnieri and Mezei for differentiating hydration propensities of different DNA grooves (Guarnieri, F. and Mezei, M., *J. Am. Chem. Soc. 118*:8493-8494 (1996)). (This reference is incorporated herein by reference in its entirety.) These results were obtained with the Metropolis Monte Carlo (MMC) code developed by the group of Mezei, Mount Sinai School of Medicine, NY. For these simulations, the system was composed of a molecule fraction of DNA surrounded by a varying number of *interacting* water molecules.

[0025] In its original form, the LMC algorithm carried out a series of calculations similar to the MMC approach for each fragment-type of interest, i.e. simulations in which both the fragment - protein as well as all fragment – fragment interactions were considered. However, it has been acknowledged that considering fragment-fragment interactions is actually detrimental to the interpretation of the simulation results for all fragments but water. Indeed, due to the high dilution of the solute molecules in actual biochemical relevant conditions, considering interactions between non-water fragments is not realistic. Furthermore, the drug leads assembled by LCD usually are composed of only one fragment of each type. Fragment-fragment interactions in the LMC simulation thus lead to undesirable correlation effects. Finally, in the original MMC code, carrying out the simulated annealing of the chemical potential for computing the free energies of binding required the data from multiple ensemble samplings at various B values. In the absence of fragment-fragment interactions however, the required data can be directly derived from the sampling of a single ensemble. As will be shown further on, this simplification results from the ability of establishing the analytical dependence in B of the fragment density when fragment interactions are omitted. This fact naturally provides an opportunity for significant computational speedup.

[0026] It turns out that the standard Metropolis Monte Carlo algorithm has difficulty in handling simulations where fragment-fragment interactions are removed. Indeed, the absence of fragment-fragment interactions leads to the possible overlap of fragments and thus to a broad range of fragment densities between the higher and lower affinity binding sites on the protein, which the standard Metropolis Monte Carlo scheme has trouble in resolving. This problem has been overcome in the current implementation of LMC by developing a weighted Metropolis Monte Carlo scheme.

[0027] The system in which fragment-fragment interactions have been removed can be referred to as being linear by reference to the linear properties of the differential equation (Liouville-type) that describes the time-evolution of the fragment density away from thermodynamic equilibrium.

II. Process

A. Formulation

[0028] First, the derivation of the single fragment density in the framework of the grand canonical ensemble is presented.

[0029] The potential energy of the system composed of *N* fragments is denoted $U(\Gamma, N)$. In general, *U* includes both contributions from fragment-protein and fragment-fragment interactions. The configuration of the system is characterized by

$$\Gamma = \left(Y_1, Y_2, \ldots, Y_N\right), \tag{1}$$

where Y_i=(x_i, Ω_i) stands for the position x_i and orientation Ω_i of fragment *i*.

[0030] In the grand canonical ensemble, the probability that the system has *N* fragments in configuration Γ is given by

$$f(\Gamma, N) = \frac{1}{Q} \frac{1}{V^N \sigma^N} \frac{1}{N!} \exp\left[BN - \beta U(\Gamma, N)\right], \tag{2}$$

with the normalization factor given by the grand partition function

$$Q = \sum_{N=0}^{\infty} \tfrac{1}{N!} \exp(BN) \int \frac{dY^N}{V^N \sigma^N} \exp\left[-\beta U(\Gamma, N)\right]. \tag{3}$$

Here V is the volume of the system, σ is the volume of orientational space, T is the temperature, $\beta = 1/K_BT$, and B is related to the excess chemical potential μ^{ex}, i.e. the energy cost in units of β^{-1} for a particle to leave the system:

$$B = \beta\mu^{ex} + \log\langle N\rangle, \tag{4}$$

where $\langle N\rangle$ is the average number of fragments in the system. The integral in Eq. (3) is taken over the whole configuration space $(V\sigma)^N$.

[0031] Assuming no fragment-fragment interactions, the potential energy U of the system becomes:

$$U(\Gamma, N) = \sum_{i=1}^{N} E(Y_i), \tag{5}$$

where $E(Y_i)$ is the energy of interaction of the ith fragment with the protein.

[0032] The grand partition function can then be written as

$$Q = \sum_{N=0}^{\infty} \tfrac{1}{N!}\left(\exp(B)\int\frac{dY}{V\sigma}\exp[-\beta E(Y)]\right)^N = \exp Z, \tag{6}$$

with

$$Z = \exp(B)\int\frac{dY}{V\sigma}\exp[-\beta E(Y)]. \tag{7}$$

[0033] In this case, the probability $P(N)$ for having N fragments in the system is given by

$$P(N) = \int dY^N f(\Gamma, N) = \exp(-Z)\frac{Z^N}{N!}. \tag{8}$$

This is simply the Poisson distribution with parameter Z. In particular, the average number of fragments in the system is given by

$$\langle N\rangle = \sum_{N=1}^{\infty} NP(N) = Z, \tag{9}$$

which thus scales exponentially with B.

[0034] In fact, more generally, the probability $P(n,\Delta V)$ of finding n fragments in any given sub-volume ΔV of configuration space is given by a Poisson distribution:

$$
\begin{aligned}
P(n,\Delta V) &\approx \sum_{N=n}^{\infty}\frac{N!}{(N-n)!n!}\int_{\Delta V} dY_1\ldots dY_n \int_{V} dY_{n+1}\ldots dY_N f(\Gamma,N) \\
&= \frac{1}{n!}\left(\exp(B)\int_{\Delta V}\frac{dY}{V\sigma}\exp[-\beta E(Y)]\right)^n \frac{1}{Q}\sum_{N=n}^{\infty}\frac{Z^{N-n}}{(N-n)!} \qquad (10)\\
&= z^n / n!,
\end{aligned}
$$

with

$$z = \exp(B)\int_{\Delta V}\frac{dY}{V\sigma}\exp[-\beta E(Y)]. \qquad (11)$$

[0035] Finally, the single fragment density is given by

$$
\begin{aligned}
f_{gc}(Y) &= \sum_{N=1}^{\infty} N\int dY_2\ldots\int dY_N f\left(\Gamma = \left(Y, Y_2, \ldots, Y_N\right), N\right) \\
&= \exp(-Z)\frac{1}{V\sigma}\exp[B-\beta E(Y)]\sum_{N=1}^{\infty}\frac{1}{(N-1)!}Z^{(N-1)} \\
&= \frac{1}{V\sigma}\exp[B-\beta E(Y)], \qquad (12)
\end{aligned}
$$

which again scales exponentially with respect to B. Here the subscript 'gc' stands for Grand Canonical.

[0036] Note that one recovers Eq. (9) for the average number of fragments in the system by integrating f_{gc} over all configurations:

$$\int dY\, f_{gc}(Y) = Z. \qquad (13)$$

B. Numerical Method

[0037] Equation (12) for the single fragment density shows the large dynamical range that may result from the exponential dependence of this quantity with respect to the single fragment-protein potential energy *E(Y)*. This dependence results from the possible overlap of the non-interacting fragments. This is not an issue in the presence of fragment-fragment interactions, as an upper bound to the fragment density is set by the tightest possible packing of the molecules.

[0038] The underlying method developed for the WGCMMC approach to enable the accurate resolution of the above-mentioned dynamical range in densities is presented here.

[0039] For numerical purposes, instead of considering a constant *B* value, one may consider a field $B_{num}(Y)$ in the single particle configuration space *Y*. This field can be interpreted as the energy cost for a particle to leave the system specifically from position *Y*. Instead of Eq. (2), the density of states in this generalized grand canonical ensemble is now given by

$$f_{\text{num}}(\Gamma, N) = \frac{1}{Q_{\text{num}}} \frac{1}{V^N \sigma^N} \frac{1}{N!} \exp\left[\sum_{i=1}^{N} B_{\text{num}}(Y_i) - \beta U(\Gamma, N)\right], \qquad (14)$$

with the normalization factor (grand partition function) now given by

$$Q_{\text{num}} = \sum_{N=0}^{\infty} \frac{1}{N!} \int \frac{dY^N}{V^N \sigma^N} \exp\left[\sum_{i=1}^{N} B_{\text{num}}(Y_i) - \beta U(\Gamma, N)\right]. \qquad (15)$$

An analogous derivation as the one used for obtaining Eq. (12) leads to the corresponding single fragment density:

$$f_{\text{gc,num}}(Y) = \frac{1}{V\sigma} \exp\left[B_{\text{num}}(Y) - \beta E(Y)\right]. \qquad (16)$$

Thanks to the field $B_{num}(Y)$, one now has a direct handle on the value of the density in each position *Y* of the single particle configuration space. Thus, by iteratively adapting $B_{num}(Y)$ during the convergence phase of the Metropolis Monte Carlo simulation, one may obtain appropriate sampling in all regions of interest. For a B_{num} field continuous over Y, this would be achieved by taking

$$B_{\text{num}}(Y) \simeq \min\left(\beta E(Y) + \text{const}, B_{\max}\right), \qquad (17)$$

leading to similar numerical densities of fragment instances in various regions of space. An upper bound B_{max} is set on B_{num} to avoid unnecessary sampling in strongly unfavorable positions, i.e., essentially for configurations leading to steric clashes. This ensures to preserve the advantages of the Metropolis Monte Carlo scheme over standard Monte Carlo integration algorithms. In practice, the field $B_{num}(Y)$ is usually chosen to be independent of the fragment orientation, and to be piece-wise constant on a 3-D grid in x-space (translational-space). Eq. (16) and (17) also show how the purpose of the

$B_{num}(Y)$ field could have equivalently been achieved by rescaling the single fragment potential energy field E(Y).

[0040] Making use of the exponential dependence in *B* of the density, one can infer the physical fragment density $f_{\mathrm{gc}}(Y)$ at any $B=B_0=$ constant value from the simulation results for a given numerical $B_{num}(Y)$ field. Assume that one has a sampling $\{\Gamma_i=(Y_1,\ldots,Y_{N_i})\}_{i=1,\ldots,n_{\mathrm{snap}}}$ of n_{snap} snapshots from the numerical distribution $f_{\mathrm{gc,num}}(\Gamma,\mathrm{N})$. The average of any single fragment quantity $A(Y)$ over the distribution $f_{\mathrm{gc}}(Y)$ is then given by

$$\begin{aligned}\langle A\rangle &= \int dY f_{\mathrm{gc}}(Y)A(Y) = \int dY f_{\mathrm{gc,num}}(Y)\frac{f_{\mathrm{gc}}(Y)}{f_{\mathrm{gc,num}}(Y)}A(Y)\\ &\simeq \frac{1}{n_{\mathrm{snap}}}\sum_{i=1}^{n_{\mathrm{snap}}}\sum_{j=1}^{N_i} w_j A(Y_j),\end{aligned} \tag{18}$$

where w_j is the weight assigned to the fragment state Y_j, and defined by

$$w_j = \frac{f_{\mathrm{gc}}(Y_j)}{f_{\mathrm{gc,num}}(Y_j)} = \exp\left(B_0 - B_{\mathrm{num}}(Y_j)\right). \tag{19}$$

Results for any B value can thus be inferred from Eqs. (18)-(19). In particular, as will be presented in more detail, by omitting fragment-fragment interactions, simulated annealing of the chemical potential (i.e. variation of B) can be derived analytically given the sampling data for a single $B_{num}(Y)$ field.

C. Handling WGCMMC data

[0041] The following addresses how the WGCMMC data is to be handled and analyzed.

[0042] The starting point for the data interpretation is the relation linking the WGCMMC data to the association constant K_a characterizing the binding of the considered fragment to a given region on the protein. This relation for K_a is rederived here.

[0043] The association constant K_a characterizes the equilibrium of the binding process

$$\mathrm{F}+\mathrm{P}\leftrightarrow \mathrm{FP}, \tag{20}$$

and is defined by

$$K_a = \frac{[\mathrm{FP}]}{[\mathrm{F}][\mathrm{P}]}, \tag{21}$$

where [P], [F], and [FP] are respectively the concentrations of protein P alone, fragment F alone, and of a particular protein-fragment complex FP (binding mode). The association constant is the basic biologically relevant quantity.

[0044] Let us consider a single protein in a volume *V*. For the sake of the following discussion, take *V* to be large, although for the actual LMC simulation this need not be the case. The protein concentration is thus given by [P] = 1/V. Furthermore, let us note *n* the average number of fragments in the binding volume ΔV_b (in general a volume with limits both in translational and orientational space), and *N* the average total number of fragments in the system, so that [F] = (N – n)/V and [FP] = n/V. The association constant can thus be written

$$K_a = \frac{n/V}{(N-n)/V\,1/V} \simeq V\frac{n}{N} \tag{22}$$

having invoked the thermodynamic limit of large volume *V*, so that $n<<N$ ($N/V \rightarrow$ const, for $V \rightarrow \infty$). The values *n* and *N* can be obtained from the fragment density (12):

$$n = \int_{\Delta V_b} dY\, f_{gc}(Y) = \frac{e^B}{V\sigma}\int_{\Delta V_b} dY \exp\left[-\beta E(Y)\right], \tag{23}$$

$$N = \int_V dY\, f_{gc}(Y) = \frac{e^B}{V\sigma}\int_V dY \exp\left[-\beta E(Y)\right] \simeq e^B, \tag{24}$$

having again invoked the assumption of the high protein dilution, so that the total system volume *V* is much larger than the effective region of interaction between the fragment and the protein, and thus one may consider $E(Y) \cong 0$ in deriving the last approximate equality in (24). The association constant now becomes:

$$K_a = \frac{1}{\sigma}\int_{\Delta V_b} dY \exp\left[-\beta E(Y)\right]. \tag{25}$$

[0045] On the basis of Eq. (25) one can also write the association constant in terms of the free energy of binding ΔA:

$$K_a = V\exp\left(-\beta\Delta A\right). \tag{26}$$

where $\Delta A = A_{\mathrm{FP}} - \mathrm{A_F}$, with A_{FP} and A_{F} respectively the free energies of the fragment-protein complex FP and of the fragment F alone:

$$A_{\mathrm{FP}} = -\frac{1}{\beta}\log\left(\int_{\Delta V_b} dY\exp\left[-\beta E(Y)\right]\right), \tag{27}$$

$$A_{\mathrm{F}} = -\frac{1}{\beta}\log\int_V dY = -\frac{1}{\beta}\log(V\sigma). \tag{28}$$

[0046] The critical value $\mathrm{B_c}$ that is associated to the binding volume $\Delta\mathrm{V_b}$ can be defined as the value for which the average number of fragments in the binding site is one. From Eq. (23) follows:

$$n\left(B_c\right) = 1 \quad \Leftrightarrow \quad e^{-B_c} = \frac{1}{V\sigma}\int_{\Delta V_b} dY\exp\left[-\beta E(Y)\right], \tag{29}$$

and from (25), (26) and (29) one sees that $\mathrm{B_c}$ is directly related to $\mathrm{K_a}$ and $\Delta\mathrm{A}$ as follows:

$$K_a = Ve^{-B_c}, \tag{30}$$

$$\Delta A = \frac{1}{\beta}B_c. \tag{31}$$

Thus, a low $\mathrm{B_c}$ value reflects a high affinity binding mode, and inversely a high $\mathrm{B_c}$ value reflects a low affinity mode.

[0047] The critical value B_c can be computed from the WGCMMC data using definition (29), as well as Eqs (18) and (19):

$$1 = n(B_c) = \int_{\Delta V_b} dY f_{gc}(Y) \cong \frac{1}{n_{snap}}\sum_{i=1}^{n_{snap}}\sum_{frag\, j\in\Delta V_b}\exp\left[B_c - B_{num}(Y_j)\right]$$

$$\Leftrightarrow \quad B_c = -\log\left(\frac{1}{n_{snap}}\sum_{i=1}^{n_{snap}}\sum_{frag\, j\in\Delta V_b}\exp\left[-B_{num}(Y_j)\right]\right). \tag{32}$$

Equations (30), (31) and (32) provide the basic relations for interpreting the WGCMMC data.

Binding Analysis

[0048] A first estimate of the binding affinity of a given fragment for different regions on the protein surface can be obtained by assigning a critical B_c to each fragment-residue pair. These B_c values are obtained from the WGCMMC data by applying relation (32), where the volume ΔV_b is approximated for each residue on the basis of the following proximity criteria: A fragment state is considered to be in proximity of a given residue if at least one fragment-protein atom pair *(a, b)* is such that

$$r_{ab} < \alpha \left(R_{VdW,a} + R_{VdW,b}\right), \qquad (33)$$

where r_{ab} is the distance between the two atoms, R_{VdW} is the Van der Walls radii defined as half the Lennard-Jones parameter from the AMBER force-field, and α is a numerical parameter (typically $\alpha = 1.2$).

[0049] The volume defined on the basis of the proximity criteria is in general only a crude estimate of a binding mode volume. The corresponding B_c values must therefore be interpreted accordingly. Nonetheless, comparing sets of $\mathrm{B_c}$ values obtained in this way for different fragments has proven valuable to help identify protein binding sites as follows: A binding site is identified as a set of neighboring residues with low $\mathrm{B_c}$ values (high affinity) for multiple fragments with different physico-chemical properties. This approach is based on the assumption that diverse interactions in a localized region are the necessary condition for ensuring the specificity of a binding site. This numerical identification of binding sites is preferably complemented by experimental binding information such as co-crystal X-ray data and mutational analysis.

[0050] More detailed calculations of the binding mode volumes ΔV_b, compared to the above described residue-based proximity criteria, are necessary to provide more accurate estimates of the free energy of binding using Eq. (32). Such improved binding mode volume estimates are determined by identifying "humps" in the fragment distribution. This can be achieved by clustering sampled fragment states belonging to a same potential energy well. For this purpose one makes use of the potential energies saved for the sampled fragment states.

Chemistry Design

[0051] With the purpose of data reduction, the LCD chemistry design software clumps the sampled fragment instances together. Clumping in LCD is usually carried out at a very fine-grained level, so that the clumping volume ΔV_c (limited both in translational and orientational space) is different from a true binding volume ΔV_b of the fragment. In fact, a binding mode volume is usually composed of many clump volumes. Each clump is assigned the B_c value of the binding mode volume to which it belongs.

[0052] Using the WGCMMC-type data, average clump positions (x_c) and quaternion representation (q_c) of average clump orientation can be computed by the following weighted averages:

$$\langle \mathrm{x_c} \rangle = \frac{\sum_i w_i \mathrm{x_i}}{\sum_i w_i}, \tag{37}$$

$$\langle q_c \rangle = \sum_i w_i q_i \quad \rightarrow \quad \text{Normalize}\, q_c, \tag{38}$$

where the sums are over all fragments i in the clump.

[0053] Within the chosen protein binding site, clumps of different fragment types can then be assembled into actual candidate drug leads, usually composed of four to five fragments. Assembly of fragments is carried out based on binding affinity of the different fragments (B_c values), and on geometric proximity using a variety of rules by which organic fragments may bond together, as is well known in the art.

D. Process Implementation

[0054] In light of the above analytical description of WGCMMC processing, the logic for WGCMMC can be implemented in the broader simulation context as illustrated in FIG. 1, according to an embodiment of the invention. The overall process starts at step 110. In step 120, a model is constructed for the molecules to be simulated, i.e., a protein and some number of rigid molecular fragments that may interact with the protein. In step 130, the thermodynamic equilibrium of the system is modeled so that the interactions between the fragments and the protein at thermodynamic equilibrium can be understood. This step results in simulation data that includes, for each fragment, the fragment's position, orientation, weight, and fragment-protein energy. In step 140, potential binding sites are identified on the protein. In step 150, fragments are assembled into drug leads. The overall process concludes at step 160. Each of these steps is described in greater detail below.

Molecule preparation

[0055] Step 120, the preparation of the molecular model, is illustrated in FIG. 2. This process starts at step 210. Protein preparation takes place in step 220. A protein can be viewed as a biological macro-molecule to which a prospective ligand binds. The basic protein structure is provided by experimental X-ray crystallography data, typically downloaded from a data base. The protein structure is completed for missing substructures, which in some cases may be a limited number of heavy atoms or, in other cases, entire segments of an amino-acid chain. Hydrogen atoms, not resolved by X-ray crystallography, are added as well. Conformer and protonation state issues for the amino-acids HIS, ASP, GLU, CYS, TYR, THR, and SER are also resolved at this stage. Such a process for protein preparation is disclosed and claimed in a co-pending U.S. patent application, Serial No. 60/450,711, filed on March 3, 2003, and incorporated herein by reference in its entirety.

[0056] Fragment preparation takes place in step 230. The structure and partial charges of the small organic fragments are completed with an *ab initio*, i.e., quantum mechanical based, code. This calculation is typically carried out in the framework of the Density Functional Theory (DFT) approximation using the code Gaussian (M. J. Fish et.al., "Gaussian 98, revision A.9," 1998. Gaussian Inc., Pittsburgh, PA). This step also assigns the AMBER types to each fragment atom. The process concludes at step 240.

[0057] The step of modeling the thermodynamic system is illustrated in greater detail in FIG. 3, according to an embodiment of the invention. The process starts at step 310. In step 320, a convergence phase of the weighted Metropolis Monte Carlo simulation is executed. This is followed by a sampling phase in step 330. Steps 320 and 330 are described in greater detail below. The resulting simulation data is saved in step 340. The process concludes in step 350.

Convergence phase of LMC simulation

[0058] Step 320, the convergence phase of the LMC simulation, is illustrated in FIG. 4. In this first stage of the LMC simulation, the numerical B-field, B_{num}, and the Markov chain generated by the LMC stepping are converged.

[0059] The process starts with step 410. Initially one starts with a constant field $B_{\text{num}} \equiv$ Bo, as shown in step 420. Through successive MC steps, fragment states are then sampled using the standard Metropolis Monte Carlo scheme for Grand Canonical simulations (Adams, D.J., *Molecular Physics 29*:307-311 (1975); Mezei, M., *Molecular Physics 61*:565-582 (1987)), incorporated herein by reference in their entireties. At regular intervals in the stepping of the convergence phase, sufficiently long to ensure decorrelation of states, the fragment distributions are monitored.

[0060] More exactly, in step 430, the simulation space is subdivided with a grid. Typically, the 3-dimensional translational space of the simulation system is subdivided by a structured, orthogonal, and equidistant grid, with centers x_i. Grid size is based on the variation scale of the interaction field, typically of the order of one Angstrom. The detail of the fragment distribution monitoring is given in step 440, where the weighted number of sampled fragments in each grid cell is computed as follows:

$$n_{B=0}(x_i) = \frac{1}{n_{\text{samples}}} \sum_{\text{samples}} \sum_{\text{frag j in cell i}} \exp\left[-B_{\text{num}}(Y_j)\right], \qquad (40)$$

where n_{samples} is the number of samples taken up to any point in the convergence phase. Equation (40) is an application of Eqs. (18)-(19) for B_0=0.

[0061] Based on these statistics, the field $B_{\text{num}}(x)$ is then adapted in step 450, by making use of the exponential dependence in $B_{\text{num}}(x)$ of the number of fragments in each grid cell *i*. In this way, each cell is assigned a constant value $B_{\text{num}}(x_i)$ as follows:

$$B_{\text{num}}(x_i) = \log\left(\frac{n_{\text{target}}}{n_{B=0}(x_i)}\right), \tag{41}$$

the goal being to achieve a similar average number of sampled fragments n_{target} within all cells. An upper bound B_{max} is set on B_{num} to avoid spending too much computing time on sampling very unfavorable positions, i.e., mainly for configurations leading to steric clashes or for fragment states far away from the protein surface where the binding interaction is low. In this way one still ensures the numerical advantages of the Metropolis Monte Carlo scheme over basic Monte Carlo integration algorithms.

[0062] Adapting the field $B_{\text{num}}(x)$ is an iterative process carried out through periodic updates. Indeed, the first B_{num} updates are based on some very non-uniform sampling, thorough in deep energy pockets, but poor in shallow ones. As the $B_{\text{num}}(x)$ field is adapted, the sampling is globally improved and the adjustment of $B_{\text{num}}(x)$ can be further refined.

[0063] In step 460 of the convergence phase, the $B_{\text{num}}(x)$ field is finally kept fixed, which enables the Markov chain to fully equilibrate.

[0064] The acceptance probabilities for the various types of Monte Carlo steps in the framework of the Grand-Canonical ensemble with spatially varying $B_{\text{num}}(x)$ field are as follows:

-- Moving a fragment within the simulation system: Assuming symmetric attempts, moving a fragment from position $Y_a = (x_a, \Omega_a)$ to position $Y_b = (x_b, \Omega_b)$ is accepted with probability:

$$\text{acc}(Y_a \rightarrow Y_b) = \min(1, \alpha), \tag{42}$$

with

$$\alpha = \exp([B(x_b) - B(x_a)] - \beta[E(Y_b) - E(Y_a)]). \tag{43}$$

-- Inserting a fragment into the simulation system: Assuming no biased sampling, such as preferential sampling or cavity bias, and considering that

N fragments are already present in the system, the probability of accepting the insertion of a fragment at position $Y = (\mathrm{x}, \Omega)$ is given by:

$$\mathrm{acc}(N \rightarrow N+1) = \min(1, \alpha), \tag{44}$$

with

$$\alpha = \frac{1}{N+1}\exp\big(B(\mathrm{x}) - \beta E(Y)\big). \tag{45}$$

-- Deleting a fragment from the simulation system: The probability of deleting a fragment at position $Y = (\mathrm{x}, \Omega)$, assuming that $N + 1$ fragments are initially in the system, is given by:

$$\mathrm{acc}(N+1 \rightarrow N) = \min(1, \alpha), \tag{46}$$

with

$$\alpha = (N+1)\exp\big(-B(\mathrm{x}) + \beta E(Y)\big). \tag{47}$$

Equations (42) to (47) can be generalized to various types of biased sampling.

Sampling phase of MC simulation

[0065] The numerical B-field, B_{num}, is kept fixed throughout the second stage, the so-called sampling phase, of the MC simulation. This phase, step 330 of FIG. 3, is illustrated in greater detail in FIG. 5. The process starts with step 510. In step 520, $B_{\mathrm{num}}(\mathrm{x})$ is kept fixed. In step 530, the equilibrated Markov chain is sampled periodically at successive decorrelated states until sufficient sampling data is acquired. In step 540, for each sampled state of the system, the positions x, orientations Ω, weights w = exp($-B_{\mathrm{num}}$(x)), and fragment-protein potential energies E(Y) of all fragments present in the system are saved. The process concludes at step 550.

Identifying binding modes

[0066] FIG. 6 illustrates the process of identifying potential binding sites, according to an embodiment of the invention. The process starts with step 610. In step 620, logic such as the Locus Binding Analysis (LBA) software package begins execution. In step 630, a value B_c is assigned to each fragment-residue pair. In step 640, potential binding sites are identified on the basis of the B_c values. As discussed above, these B_c values are obtained from the WGCMMC data by applying relation (32), where the volume ΔV_b is defined for each residue on the basis of the proximity criteria. Recall from Eq. (33) above that a fragment is considered to be in proximity of a given residue if at least one fragment-protein atom pair *(a, b)* is such that

$$r_{ab} < \alpha \left(R_{VdW,a} + R_{VdW,b}\right), \tag{33}$$

where r_{ab} is the distance between the two atoms, R_{VdW} is the Van der Walls radii defined as half the Lennard-Jones parameter from the AMBER force-field, and typically $\alpha = 1.2$. A binding site is then identified as a set of residues with low B_c values (high affinity) for multiple fragments with diverse physico-chemical properties. The process concludes at step 650.

Assembling fragments in the binding site

[0067] Step 150 of FIG. 1, the step of assembling fragments into drug leads, is illustrated in greater detail in FIG. 7, according to an embodiment of the invention. The process starts with step 710. With the purpose of data reduction, fragment instances are clumped together in step 720. Clumping is carried out at a very fine-grained level (both in translational and orientational space), so that the clumping volume ΔV_c is different from a true binding volume. In fact, a binding mode volume is usually composed of many clump volumes. The purpose of this clumping is to achieve some level of data reduction before carrying on with the fragment assembly into drug leads, involving computationally labor intensive combinatorics.

[0068] In steps 730 through 750, weighted average clump positions (x_c) and quaternion representation of weighted average clump orientation (q_c) is computed as described earlier:

$$\langle \mathrm{x_c} \rangle = \frac{\sum_i w_i \mathrm{x_i}}{\sum_i w_i}, \tag{37}$$

$$\langle q_c \rangle = \sum_i w_i q_i \quad \rightarrow \quad \text{Normalize}\, q_c, \tag{38}$$

where the sums are over all fragments *i* in the clump.

In the same way, as appears in step 760, one may also compute the average potential energy of the clump:

$$\langle E_c \rangle = \frac{\sum_i w_i E_i}{\sum_i w_i}, \qquad (39)$$

where E_i is the potential energy of interaction of fragment i with the protein.

[0069] In step 770, each clump is assigned the B_c value of the binding mode volume to which it belongs.

[0070] In step 780, within the chosen protein binding site, clumps of different fragment types are then assembled into actual candidate drug leads, usually (though not always) composed of four to five fragments. Assembly of fragments is carried out based on binding affinity of the different fragments (B_c values), and on geometric proximity, using a variety of rules by which organic fragments may bond together as is well known in the art.

III. Computing Environment

[0071] The present invention may be implemented using software and may be implemented in conjunction with a computing system or other processing system. An example of such a computer system 800 is shown in FIG. 8. The computer system 800 includes one or more processors, such as processor 804. It is to be noted that the here-described fragment-based computation is particularly well suited for being carried out on a computer cluster, each cluster node computing the interaction of a given fragment type with the target protein. The processor 804 is connected to a communication infrastructure 806, such as a bus or network. Various software implementations are described in terms of this exemplary computer system. After reading this description, it will become apparent to a person skilled in the relevant art how to implement the invention using other computer systems and/or computer architectures.

[0072] Computer system 800 also includes a main memory 808, preferably random access memory (RAM), and may also include a secondary memory

810. The secondary memory 810 may include, for example, a hard disk drive 812 and/or a removable storage drive 814, representing a magnetic tape drive, an optical disk drive, etc. The removable storage drive 814 reads from and/or writes to a removable storage unit 818 in a well-known manner. Removable storage unit 818 represents a magnetic tape, optical disk, or other storage medium that is read by and written to by removable storage drive 814. As will be appreciated, the removable storage unit 818 can include a computer usable storage medium having stored therein computer software and/or data.

[0073] In alternative implementations, secondary memory 810 may include other means for allowing computer programs or other instructions to be loaded into computer system 800. Such means may include, for example, a removable storage unit 822 and an interface 820. An example of such means may include a removable memory chip (such as an EPROM, or PROM) and associated socket, or other removable storage units 822 and interfaces 820 which allow software and data to be transferred from the removable storage unit 822 to computer system 800.

[0074] Computer system 800 may also include one or more communications interfaces, such as network interface 824. Network interface 824 allows software and data to be transferred between computer system 800 and external devices. Examples of network interface 824 may include a modem, a network interface (such as an Ethernet card), a communications port, a PCMCIA slot and card, etc. Software and data transferred via network interface 824 are in the form of signals 828 which may be electronic, electromagnetic, optical or other signals capable of being received by network interface 824. These signals 828 are provided to network interface 824 via a communications path (i.e., channel) 826. This channel 826 carries signals 828 and may be implemented using wire or cable, fiber optics, an RF link and other communications channels.

[0075] In this document, the terms "computer program medium" and "computer usable medium" are used to generally refer to media such as removable storage units 818 and 822, a hard disk installed in hard disk drive

812, and signals 828. These computer program products are means for providing software to computer system 800.

[0076] Computer programs (also called computer control logic) are stored in main memory 808 and/or secondary memory 810. Computer programs may also be received via communications interface 824. Such computer programs, when executed, enable the computer system 800 to implement the present invention as discussed herein. In particular, the computer programs, when executed, enable the processor 804 to implement the present invention. Accordingly, such computer programs represent controllers of the computer system 800. Where the invention is implemented using software, the software may be stored in a computer program product and loaded into computer system 800 using removable storage drive 814, hard drive 812 or communications interface 824.

IV. Conclusion

[0077] While various embodiments of the present invention have been described above, it should be understood that they have been presented by way of example, and not limitation. It will be apparent to persons skilled in the relevant art that various changes in detail can be made therein without departing from the spirit and scope of the invention. Thus the present invention should not be limited by any of the above-described exemplary embodiments.